A joint scoring model for peer-to-peer and traditional lending

A joint scoring model for peer-to-peer and traditional lending: a bivariate model with copula dependence

Raffaella Calabrese
Credit Research Centre and Business School, University of Edinburgh, Edinburgh, UK.
E-mail: raffaella.calabrese@ed.ac.uk
Silvia Angela Osmetti
Department of Statistical Science, Universitá Cattolica del Sacro Cuore, Milan, Italy.
E-mail: silvia.osmetti@unicatt.it
Luca Zanin
Wealth and Asset Management, Prometeia, Bologna, Italy.
E-mail: luca.zanin@studio.unibo.it

Abstract.
We analyse the dependence between defaults in peer-to-peer (P2P) lending and credit bureaus. To achieve this aim, we propose a new flexible bivariate regression model suitable for binary imbalanced samples. We use different copula functions to model the dependence structure between defaults in the two credit markets. We implement the model in the `R` package `BivGEV` **and we explore the empirical properties of the proposed fitting procedure by a Monte Carlo study.** The application of this proposal to a comprehensive dataset provided by Lending Club shows a significant level of dependence between the defaults in P2P and credit bureaus. Finally, we find that our model outperforms **the bivariate probit and univariate logit** in predicting P2P default, in estimating the Value at Risk and the Expected Shortfall.

Keywords: Binary imbalanced samples; Copula based model; Credit bureau; Generalised extreme value regression model; Peer-to-peer lending

1. Introduction

Peer-to-peer (P2P) lending allows direct lending between lenders and borrowers using a platform that acts as a broker between them, without involving traditional financial institutions, such as banks. On the platform, borrowers submit their requests for loan amounts and lenders are able to fund these requests. Each borrower's request is usually funded by multiple lenders. Transaction fees are usually charged at origination so that platforms can make profits. In recent years, the P2P lending market has shown a substantial increase in popularity (Bachmann et al., 2011; Berger and Gleisner, 2009; Lin et al., 2017; Milne and Parboteeah, 2016; Wang et al., 2009). According to a PriceWaterhouseCoopers report (PWC, 2015) in 2014 P2P lending generated approximately \$5.5 billion worth of loans in the US. PwC estimates that the market could reach \$150 billion or higher by 2025.

Matching the supply and demand of funds through an online platform can generate information asymmetry between lenders and borrowers, as the creditworthiness of borrowers is unknown to lenders. The main consequences of such asymmetric information can be moral hazard and adverse selection (Stiglitz and Weiss, 1981). High street banks provide credit only after an extensive check of the borrowers' financial conditions, usually employing credit scoring models. This increases the level of trust that lenders have towards borrowers and is likely to reduce adverse selection and moral hazard. However, it is more difficult to reproduce this confidence in the online environment as there isn't any direct relationship between lenders and borrowers. The main aim of this paper is to improve the assessment of default risk in P2P lending by gaining a better understanding of the relationship between the defaults of P2P and credit bureaus.

The last financial crisis has decreased the availability of credit to different kinds of borrowers in various regions from 2007-2008 for few years. The availability of household unsecured credit started increasing again only in 2010 in the UK (Bank of England, 2018). In the US, the census shows that the percentage of households holding some form of debt decreased from 74% in 2000 to 69% in 2011. Also bank lending to firms declined by about 9% from 2008 to 2011 in the US, where small businesses have been far more severely affected than large firms with a decline in bank lending of almost 18% (Cole, 2012). The provision of credit by banks to small and medium enterprises has been substantially reduced also in the Euro area (ECB, 2018) and in the UK (Zhao and Jones-Evans, 2016).

As a result of this credit contraction, alternative financial services have been developed in the financial technology industry (fintech), such as P2P lending platforms. Before providing credit, such online platforms need to estimate a scoring model that discriminates between potential good and bad borrowers. To improve the predictive accuracy of a scoring model for P2P lending, we suggest to estimate the P2P default probability conditional on a loan being in default or not in credit bureaus. To the best of our knowledge, this is the first paper that analyses the dependence between P2P and bank loan defaults and that uses this dependence to improve the predictive accuracy of a scoring model for P2P lending.

To perform this analysis, we propose a flexible bivariate regression model for binary imbalanced outcomes. We consider two binary dependent variables, one represents if the borrower is in default in the P2P online platform and the other represents if the same borrower is listed as being in default by a credit bureau. A common characteristic of empirical studies on default risk in both the P2P lending (e.g. Serrano-Cinca et al., 2015) and the traditional banking market (e.g. Mian and Sufi, 2013) is to obtain a percentage of bad loans that are substantially lower than that of good loans, so the binary sample is defined as being imbalanced. Previous studies (Andreeva et al., 2016; Calabrese and Osmetti, 2013; Wang and Dey, 2010) show in the univariate context that the use of a symmetric link function, such as the logit or probit link, may not be appropriate for imbalanced samples. Furthermore, the maximum likelihood estimators of the regression parameters in a logistic model can be biased if the binary sample is imbalanced (King and Zeng, 2001).

As the characteristics of bad borrowers are more informative than those of good borrowers, Calabrese and Osmetti (2013) suggested to use an asymmetric link function

that assign more importance to the information on defaulted loans. As the aim of a scoring model is to estimate the probability of default, bad borrowers represent the right tail of the response curve in a binary regression model. The Generalised Extreme Value (GEV) distribution is a flexible function used in the literature to analyse the tail of a distribution (Dey and Yan, 2016). Therefore, we suggest to use the GEV link function to model the marginal probabilities in a bivariate regression. To study the dependence between the marginal default probabilities we use a copula approach (Nelsen, 2006) for its flexibility, analogously to Genest et al. (2013) and Radice et al. (2016). We call this proposal the Bivariate Generalised Extreme Value model (BivGEV). We apply the maximum likelihood method to estimate the parameters of the BivGEV model and we implement **it in the R environment using the package** `BivGEV` (see the supplementary material).

We use the BivGEV model to analyse 18,113 P2P loans from 2010 to 2012 provided by Lending Club, the biggest US P2P lending platform. There are few recent studies that analyse the determinants of default in the P2P platforms, for example Guo et al. (2016), Dorfleitner et al. (2016), Lin et al. (2017), Serrano-Cinca et al. (2015). Knowing if a borrower has a defaulted payment in the last 7 years, we can first analyse the dependence between P2P and bank loan defaults. As the information on P2P default is sequential to that from credit bureaus, P2P platforms usually use **univariate logits** to build scoring models that consider as explanatory variable if the borrower has been listed in default or not in credit bureaus (Lin et al., 2017). Instead, we suggest to discriminate between good and bad borrowers in the P2P platform using the P2P default probability estimated conditional on a loan being in default or not in credit bureaus. The empirical analysis shows that our proposal provides a superior performance in predicting defaults and a more accurate estimate of the Value at Risk and the Expected Shortfall compared to those obtained by **univariate logits**, traditionally used in industry and in academic research.

The paper is organised as follows. In Section 2, we present the BivGEV model for binary imbalanced response variables and the estimation procedure. **In Section 3 we perform a Monte Carlo study.** In Section 4, we analyse the dependence between defaults in P2P lending and credit bureaus using data from Lending Club. We show the best predictive accuracy of our models compared to the models used in industry. Finally, section 6 is devoted to the concluding remarks.

2. The bivariate generalised extreme value regression for binary imbalanced responses

2.1. The univariate GEV model for the marginal defaults

We consider a portfolio of n loans, the binary outcomes y_i

$$y_i = \begin{cases} 1 & if\ the\ borrower\ i\ defaults \\ 0 & otherwise \end{cases}$$

and the p covariates $x_{i1}, x_{i2}, ..., x_{ip}$, with $i = 1, 2, ..., n$. The most used models to estimate the default probability $\pi_i = P(Y_i = 1|x_{i1}, x_{i2}, ..., x_{ip})$ are the logistic and the probit models. When the binary dependent variable Y is rare, for example in a credit

portfolio, the logistic and probit models have some drawbacks (King and Zeng, 2001 and Calabrese and Osmetti, 2013). As a symmetric link function is used in these models, the response curve approaches zero at the same rate that it approaches one. Instead, the characteristics of the rare events, represented by the defaulted borrowers $y_i = 1$, are more informative than those of the non-defaulters $y_i = 0$, so the default probability for the actual defaults is underestimated (Andreeva et al., 2016 and Calabrese et al., 2016). Previous studies (Calabrese and Osmetti, 2015 and Calabrese et al., 2016) show these disadvantages for different default percentages (1%, 2%, 5% and 10%).

To overcome this drawback and to focus the attention on the tail of the response curve for values close to 1 that represent defaults, we use an asymmetric link function. Since the Generalised Extreme Value (GEV) random variable is used in the literature for modelling the tail of a distribution (see Falk et al., 2010 and Kotz and Nadarajah, 2000), Calabrese and Osmetti (2013) suggested its quantile function as the link function in a Generalised Linear Model (GLM)

$$\frac{[-ln(\pi_i)]^{-\tau} - 1}{\tau} = \mathbf{x}_i'\boldsymbol{\beta} \;\; with \;\; \tau \in R, \tag{1}$$

where $\mathbf{x}' = (1, x_1, .., x_p)$ is the vector of the explanatory variables for the loan i and $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)'$ is the vector of the constant and the regressor parameters. The log-log and the complementary log-log link functions are asymmetric link functions used in the literature (Agresti, 2002) and they represent particular cases of the GEV model. By applying the inverse of the link function in equation (1), the response curve of the GEV model results

$$\pi(\mathbf{x}_i; \boldsymbol{\beta}, \tau) = exp\left\{ - \left[1 + \tau(\mathbf{x}_i'\boldsymbol{\beta})\right]^{-1/\tau} \right\}. \tag{2}$$

The shape parameter τ of the GEV distribution controls the tail behaviour. For different values of the parameter τ, three families of distributions are defined:

- for $\tau \to 0$, the GEV distribution is known as the Gumbel class;
- for $\tau > 0$, the GEV distribution is known as Fréchet;
- for $\tau < 0$, the GEV distribution is known as Weibull.

Several studies (Andreeva et al., 2016; Calabrese and Osmetti, 2013 and 2015; Calabrese et al., 2016; Calabrese and Giudici, 2015) show that this model is very flexible and outperforms the logit and the probit models in predicting defaults. Moreover, if the sample is imbalanced with a low percentage of $y = 1$, as in credit scoring models, the best link function is obtained for $\tau < 0$.

To extend the GEV model to the bivariate case, we use the copula function described in the following section.

2.2. The copula function

Every bivariate cumulative distribution function (cdf) $F(\cdot)$ can be considered as the result of two components: two marginal distributions and a dependence structure. A

bivariate copula function $C_\lambda : I^2 \to I$, with $I^2 = [0,1] \times [0,1]$ and $I = [0,1]$, describes the way in which the marginals cdfs $F_1(\cdot)$ and $F_2(\cdot)$ are linked together. It is the bivariate cdf of a bivariate random variable (U, V)

$$C_\lambda(u,v) = P(U \leq u, V \leq v), \quad 0 \leq u \leq 1 \quad 0 \leq v \leq 1 \tag{3}$$

where the marginal distributions of U and V are uniform over $[0,1]$ and the copula parameter $\lambda \in \Lambda$ describes the intensity of the association between the marginal random variables U and V.

The Sklar's theorem (Sklar, 1959) states that there is a function $C_\lambda : I^2 \to I$ such that

$$F_{X,Y}(x,y) = C_\lambda(F_X(x), F_Y(y)) \tag{4}$$

where (X, Y) is a bivariate random variable with joint cdf $F_{X,Y}(x,y)$ and marginals cdf $F_X(x)$ and $F_Y(y)$. If the marginals cdf are continuous then the copula is unique. Otherwise, if $F_X(x)$ and $F_Y(y)$ are not continuous, the copula $C_\lambda(\cdot,\cdot)$ is uniquely determined on $RanF_X \times RanF_Y$. Conversely, if $C_\lambda(\cdot)$ is a copula and $F_X(x)$ and $F_Y(y)$ are marginal cdfs, then the $F_{X,Y}(x,y)$ is a cdf.

Analogously, we can define the survival copula function $\widehat{C}_\lambda : I^2 \to I$ as

$$\widehat{C}_\lambda(\overline{F}_X(x), \overline{F}_Y(y)) = P(X > x, Y > y) = \overline{F}_{X,Y}(x,y), \tag{5}$$

where $\overline{F}_X(x)$ and $\overline{F}_Y(y)$ are the marginal survival cdfs and $\overline{F}_{X,Y}(x,y)$ is the bivariate survival cdf. The relationship between the copula and the survival copula is given by the following equation

$$C_\lambda(u,v) = u + v - 1 + \widehat{C}_\lambda(1-u, 1-v) \tag{6}$$

We use the Kendall's Tau coefficient to measure the association between the marginal cdfs, assuming values over the interval $[-1,1]$. We can compute it from the copula parameter λ as follows

$$Kendall - Tau = 4 \int_{I^2} C_\lambda(u,v) dC_\lambda(u,v) - 1. \tag{7}$$

Another important aspect of a copula function is the tail dependence, which measures the association between the marginal cdfs $\overline{F}_X(x)$ and $\overline{F}_Y(y)$ in the tails (Trivedi and Zimmer, 2007). Particularly, the parameter used to measure the upper tail dependence is defined as

$$\chi_u = \lim_{u \to 1^-} P[Y > F_Y^{-1}(u) | X > F_X^{-1}(u)]. \tag{8}$$

Similarly, the lower tail dependence parameter is given by

$$\chi_l = \lim_{u \to 0^+} P[Y \leq F_Y^{-1}(u) | X \leq F_X^{-1}(u)]. \tag{9}$$

Both the parameters χ_u and χ_l assume values over the interval $(0,1]$ where higher the value of the parameter, the higher the intensity of the tail dependence.

Table 1. Some characteristics of the main copula functions

Copula	Dependence	Tail Dependence	$C_\lambda(u,v)$
Gaussian	radially symmetric	no asymptotic tail dependence	$\Phi_\lambda(\Phi_1^{-1}(u),\Phi_2^{-1}(v))$ with $\lambda \in (-1,1)$
Clayton	asymmetric (exchangeable)	strong left (lower) tail dependence for $\lambda > 0$	$\max\left[(u^{-\lambda}+v^{-\lambda}-1),0\right]$ with $\lambda \in [-1,\infty$
Gumbel	asymmetric (exchangeable)	strong right (upper) tail dependence	$\exp\left(-\left[(-\ln(u))^\lambda+(-\ln(v))^\lambda\right]^{1/\lambda}\right)$ with λ
Frank	radially symmetric	no asymptotic tail dependence	$-\frac{1}{\lambda}\ln\left(1+\frac{(\exp(-\lambda u)-1)(\exp(-\lambda v)-1)}{(\exp(-\lambda)-1)}\right)$ with $\lambda \in$
Joe	asymmetric (exchangeable)	strong right (upper) tail dependence	$1-\left[(1-u)^\lambda+(1-v)^\lambda-(1-u)^\lambda(1-v)\right.$ with $\lambda \in [1,\infty)$

The widely used copula functions are the Gaussian, Clayton, Gumbel, Frank and Joe copulas, described in Table 1 (see Nelsen (2006) and Joe (1997) for details). The Gaussian and Frank copulas have radial symmetry. The Gaussian copula is the copula of the bivariate normal distribution. The parameter $-1 < \lambda_G < 1$ of the Gaussian copula represents the linear correlation coefficient. Assuming a Gaussian copula, the marginal probabilities have the same level of dependence (positive or negative) below and above their mean and there is no higher association among extreme values. On the contrary, the Clayton, Gumbel and Joe copulas are asymmetric and show tail dependence. The Clayton copula shows a strong lower tail dependence, where small values of the two marginal probabilities are more associated. **In this case, the intensity of the lower tail dependence is a function of the copula parameter** $\chi_l = 2^{-1/\lambda_{CL}}$**.**

The Gumbel copula is asymmetric (exchangeable) with strong right (upper) tail dependence, indicating that high values of the two marginal probabilities are more associated. Its parameter $\lambda_{GU} \geq 1$ is a measure of positive association and represents the intensity of the upper tail dependence ($\chi_u = 2 - 2^{1/\lambda_{GU}}$). Frank copula is a symmetric copula with weak tail dependence. The Frank copula shows positive dependence for $\lambda_F \in (0, +\infty)$, negative dependence for $\lambda_F \in (-\infty, 0)$ and independence for $\lambda_F \to 0$. Finally, Joe copula with parameter $\lambda_J \geq 1$ shows positive association and upper tail dependence ($\chi_u = 2 - 2^{1/\lambda_J}$). Figures 1 and 2 show the contour plots of the copula functions for different values of λ, corresponding to a Kendall's Tau coefficient close to 0.2, 0.5 and 0.8.

2.3. The BivGEV model

Let $\mathbf{Y} = (Y_1, Y_2)$ be a binary bivariate response variable which can assume the values $(0,0);(0,1);(1,0);(1,1)$. Y_1 describes if a loan is in default or not for credit bureaus and Y_2 if the same loan is in default or not on the P2P lending platform. We model the marginal default probabilities

$$\pi_1(\mathbf{x};\boldsymbol{\beta}_1,\tau_1) = P(Y_1 = 1|\mathbf{x};\boldsymbol{\beta}_1,\tau_1) \tag{10}$$

$$\pi_2(\mathbf{x};\boldsymbol{\beta}_2,\tau_2) = P(Y_2 = 1|\mathbf{x};\boldsymbol{\beta}_2,\tau_2) \tag{11}$$

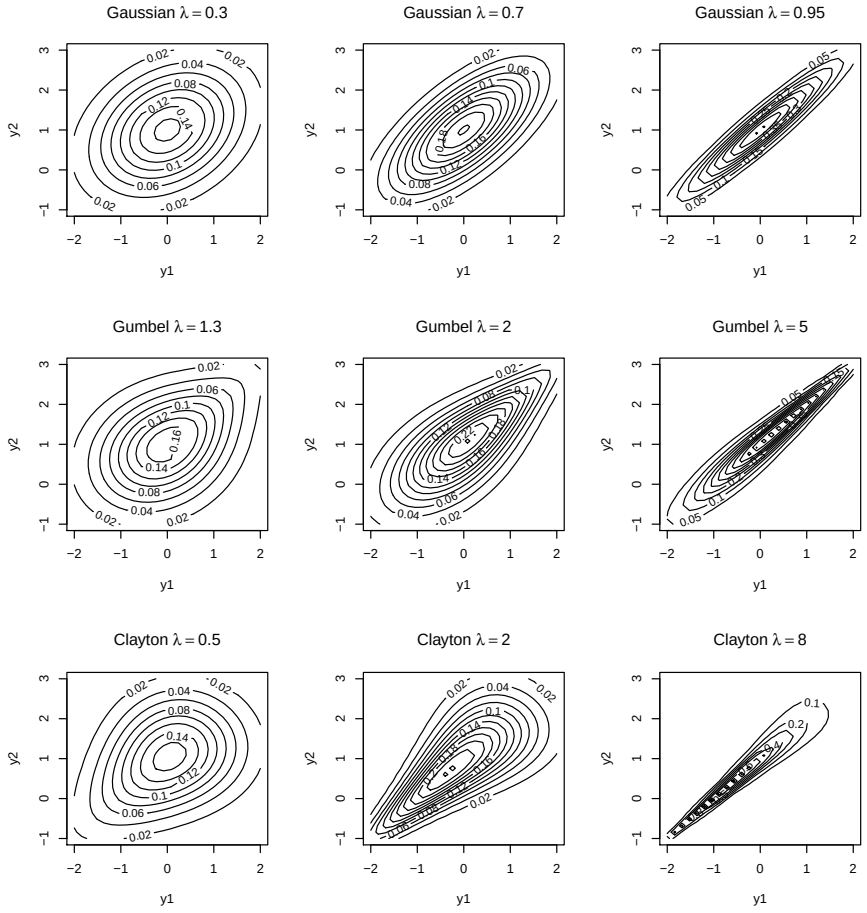

Figure 1. Contour plots of Gaussian, Gumbel and Clayton copula for several values of λ.

using the GEV model defined in equation (2).

To analyse the dependence structure between the marginal default probabilities we use a copula approach for its simplicity and flexibility (Nelsen, 2006 and Fisher, 1997). Let C_λ be the copula function defined in the equation (3) that describes the dependence between the default probabilities: $\pi_1(\mathbf{x};\boldsymbol{\beta}_1,\tau_1)$ and $\pi_2(\mathbf{x};\boldsymbol{\beta}_2,\tau_2)$. We define the joint default probability $\pi_{11}(\mathbf{x};\boldsymbol{\delta},\boldsymbol{\tau})$ as

$$\begin{aligned}\pi_{11}(\mathbf{x};\boldsymbol{\delta},\boldsymbol{\tau}) &= P(Y_1=1,Y_2=1|\mathbf{x};\boldsymbol{\delta},\boldsymbol{\tau}) \\ &= C_\lambda(\pi_1(\mathbf{x};\boldsymbol{\beta}_1,\tau_1),\pi_2(\mathbf{x};\boldsymbol{\beta}_2,\tau_2)) \\ &= C_\lambda\left(exp\left\{-\left[1+\tau_1\mathbf{x}^T\boldsymbol{\beta}_1)\right]^{-1/\tau_1}\right\}, exp\left\{-\left[1+\tau_2\mathbf{x}^T\boldsymbol{\beta}_2)\right]^{-1/\tau_2}\right\}\right)\end{aligned} \tag{12}$$

with $\boldsymbol{\delta}=(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\lambda)$ and $\boldsymbol{\tau}=(\tau_1,\tau_2)$. Therefore, the joint probability of the bivariate variable is

$$\pi_{y_1,y_2}(\mathbf{x};\boldsymbol{\delta},\boldsymbol{\tau}) = P(Y_1=y_1,Y_2=y_2|\mathbf{x};\boldsymbol{\delta},\boldsymbol{\tau})$$

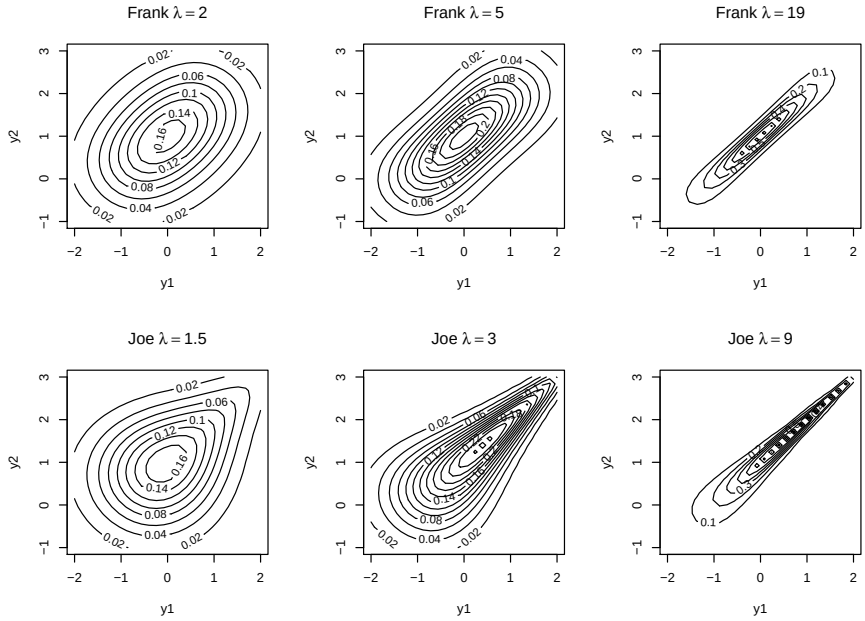

Figure 2. Contour plots of Frank and Joe copula for several values of λ.

such that

$$(Y_1, Y_2) = \begin{cases} (0,0) & \pi_{00}(\mathbf{x}; \boldsymbol{\delta}, \boldsymbol{\tau}) \\ (0,1) & \pi_{01}(\mathbf{x}; \boldsymbol{\delta}, \boldsymbol{\tau}) \\ (1,0) & \pi_{10}(\mathbf{x}; \boldsymbol{\delta}, \boldsymbol{\tau}) \\ (1,1) & \pi_{11}(\mathbf{x}; \boldsymbol{\delta}, \boldsymbol{\tau}) \end{cases}$$

where

$$\pi_{10}(\mathbf{x}; \boldsymbol{\delta}, \boldsymbol{\tau}) = \pi_1(\mathbf{x}; \boldsymbol{\beta}_1, \tau_1) - \pi_{11}(\mathbf{x}; \boldsymbol{\delta}, \boldsymbol{\tau}),$$

$$\pi_{01}(\mathbf{x}; \boldsymbol{\delta}, \boldsymbol{\tau}) = \pi_2(\mathbf{x}; \boldsymbol{\beta}_2, \tau_2) - \pi_{11}(\mathbf{x}; \boldsymbol{\delta}, \boldsymbol{\tau}),$$

$$\pi_{00}(\mathbf{x}; \boldsymbol{\delta}, \boldsymbol{\tau}) = 1 - \pi_{11}(\mathbf{x}; \boldsymbol{\delta}, \boldsymbol{\tau}) - \pi_{01}(\mathbf{x}; \boldsymbol{\delta}, \boldsymbol{\tau}) - \pi_{10}(\mathbf{x}; \boldsymbol{\delta}, \boldsymbol{\tau})$$

and $\pi_{11}(\mathbf{x}; \boldsymbol{\delta}, \boldsymbol{\tau})$ is defined by equation (12). We so propose the Bivariate GEV (BivGEV) model. For simplicity, we omit $\boldsymbol{\delta}$, $\boldsymbol{\tau}$, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and $\mathbf{x}$ from the arguments of the functions.

Each binary dependent variable Y can be represented using a latent variable Y^* (Greene, 2012) as follows

$$Y_j = \begin{cases} 1 & if \quad y_j^* > 0 \\ 0 & if \quad y_j^* \leq 0 \end{cases} \tag{13}$$

for $j = 1, 2$, where

$$Y_1^* = \mathbf{x}_1' \boldsymbol{\beta}_1 + \epsilon_1$$

$$Y_2^* = \mathbf{x}_2' \boldsymbol{\beta}_2 + \epsilon_2.$$

The error terms ϵ_1 and ϵ_2 of the two equations can be dependent in a bivariate regression model. Their dependence is described by the survival copula $\widehat{C}_\lambda$, as the following theorem

explains the relationship between the model described in equation (12) and its latent representation in equation (13).

THEOREM 2.1. *If ϵ_1 and ϵ_2 are continuous variables with marginal cdf $F_{\epsilon_j}(\varepsilon) = 1 - H(-\varepsilon)$ for $j = 1, 2$, where $H(\varepsilon) = exp\{1 - (1 + \tau\varepsilon)^{(-1/\tau)}\}$ is the extreme value cdf, and their dependence is modelled by the copula $\widehat{C}_\lambda$ such that*

$$P(\epsilon_1 \leq \varepsilon_1, \epsilon_2 \leq \varepsilon_2) = \widehat{C}_\lambda(1 - H(-\varepsilon_1), 1 - H(-\varepsilon_2)),$$

then *the bivariate regression model for (Y_1, Y_2) is given by the equation (12).*

PROOF. Let π_{11} be the joint default probability

$$\begin{aligned}
\pi_{11} &= P(Y_1^* > 0, Y_2^* > 0) = P(\epsilon_1 > -\boldsymbol{x}_1'\boldsymbol{\beta}_1, \epsilon_2 > -\boldsymbol{x}_2'\boldsymbol{\beta}_2) = \\
&= 1 - P(\epsilon_1 < -\boldsymbol{x}_1'\boldsymbol{\beta}_1, \epsilon_2 < -\boldsymbol{x}_2'\boldsymbol{\beta}_2) - P(\epsilon_1 < -\boldsymbol{x}_1'\boldsymbol{\beta}_1, \epsilon_2 > -\boldsymbol{x}_2'\boldsymbol{\beta}_2) \\
&- P(\epsilon_1 > -\boldsymbol{x}_1'\boldsymbol{\beta}_1, \epsilon_2 < -\boldsymbol{x}_2'\boldsymbol{\beta}_2) \\
&= 1 - P(\epsilon_1 < -\boldsymbol{x}_1'\boldsymbol{\beta}_1, \epsilon_2 < -\boldsymbol{x}_2'\boldsymbol{\beta}_2) - [P(\epsilon_1 > -\boldsymbol{x}_1'\boldsymbol{\beta}_1) - \pi_{11}] - [P(\epsilon_2 > -\boldsymbol{x}_2'\boldsymbol{\beta}_2) - \pi_{11}]
\end{aligned}$$

Using the survival copula function $\widehat{C}$ defined in (5), the previous equation becomes

$$\pi_{11} = \widehat{C}_\lambda(F_{\epsilon_1}(-\boldsymbol{x}_1'\boldsymbol{\beta}_1), F_{\epsilon_2}(-\boldsymbol{x}_2'\boldsymbol{\beta}_2)) + \overline{F}_{\epsilon_1}(-\boldsymbol{x}_1'\boldsymbol{\beta}_1) + \overline{F}_{\epsilon_2}(-\boldsymbol{x}_2'\boldsymbol{\beta}_2) - 1$$

where $F_{\epsilon_j}(-\boldsymbol{x}_j'\boldsymbol{\beta}_j) = P(\epsilon_j \leq -\boldsymbol{x}_j'\boldsymbol{\beta}_j)$ and $\overline{F}_{\epsilon_j}(-\boldsymbol{x}_j'\boldsymbol{\beta}_j) = P(\epsilon_j > -\boldsymbol{x}_j'\boldsymbol{\beta}_j)$.

Let u and v be the marginal default probabilities

$$u = \overline{F}_{\epsilon_1}(-\boldsymbol{x}_1'\boldsymbol{\beta}_1) = P(\epsilon_1 > -\boldsymbol{x}_1'\boldsymbol{\beta}_1) = P(Y_1^* > 0) = \pi_1$$

$$v = \overline{F}_{\epsilon_2}(-\boldsymbol{x}_2'\boldsymbol{\beta}_2) = P(\epsilon_2 > -\boldsymbol{x}_2'\boldsymbol{\beta}_2) = P(Y_2^* > 0) = \pi_2.$$

Using the equation (6), the joint default probability can be written as

$$\pi_{11} = P(Y_1 = 1, Y_2 = 1) = C_\lambda(\pi_1, \pi_2),$$

that corresponds to the equation (12).

Note that, since the errors ϵ_1 and ϵ_2 are continuous, the copula $\widehat{C}_\lambda$ between the errors is uniquely defined for the Sklar theorem.

The BivGEV model proposed in equation (12) is suitable to explain the determinants of two joint binary events in imbalanced samples. On the one hand, the flexible link function used in the BivGEV model could accommodate samples with different percentages of default (for example 5% or 1%). On the other hand, a broad class of copula functions (the most used copulas are described in Table 1) could be used to model the dependence structure between the marginal default probabilities.

2.4. The estimation procedure

For fixed values of $\boldsymbol{\tau} = (\tau_1, \tau_2)$ and **for a copula** C, given a sample of n observations the BivGEV model is estimated by maximising the complete log-likelihood function

$$l(\boldsymbol{\delta}, y_1, y_2) = \sum_{i=1}^{n} y_{1i}y_{2i}\ln(\pi_{11i}) + y_{1i}(1-y_{2i})\ln(\pi_{10i}) + y_{2i}(1-y_{1i})\ln(\pi_{01i}) + (1-y_{1i})(1-y_{2i})\ln(\pi_{00i}), \quad (14)$$

where $\boldsymbol{\delta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \lambda)$, that is

$$\begin{aligned} l(\boldsymbol{\delta}, y_1, y_2) = & \sum_{i=1}^{n} y_{i1}y_{i2}\ln[C_\lambda(\pi_{1i}, \pi_{2i})] \\ + & \; y_{i1}(1-y_{i2})\ln[\pi_{1i} - C_\lambda(\pi_{1i}, \pi_{2i})] + y_{i2}(1-y_{i1})\ln[\pi_{2i} - C_\lambda(\pi_{1i}, \pi_{2i})] \\ & + (1-y_{i1})(1-y_{i2})\ln\{1 - [\pi_{1i} + \pi_{2i} - C_\lambda(\pi_{1i}, \pi_{2i})]\}. \end{aligned}$$

We could estimate jointly the parameters $\boldsymbol{\tau}$ and $\lambda, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ by maximising the equation (14). Since the support of the joint probability density function depends on the unknown parameters τ_1 and τ_2, the estimation process would be difficult and the classical regularity conditions for maximum likelihood estimation could not be satisfied (see Smith, 1989 for details). Therefore, we fix the values of the parameters τ_1, τ_2 and we fit few BivGEV models with different values of τ_1 and τ_2. Then, we select the model with the highest predictive accuracy, analogously to Calabrese et al. (2016).

For a fixed copula function, we propose the following procedure to estimate the BivGEV model and to select the parameters τ_1 and τ_2:

(a) We specify the BivGEV model:
- We set T different values of (τ_1, τ_2) taking into account the three distribution families of the GEV model defined in Section 2.1 for $\tau_j \to 0$, $\tau_j > 0$ and $\tau_j < 0$ with $j = 1, 2$.
- **We set the dependence structure of the model choosing a copula function**.

(b) **We estimate the vector parameter $\boldsymbol{\delta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \lambda)$ for all the T BivGEV models defined in the step (a) by maximising the equation (14)**.

(c) We choose the values of the parameters (τ_1, τ_2) that minimise the MSE_+ as follows

$$MSE_+ = \frac{1}{n_1}\sum_{i=1}^{n_1}(1 - \pi_{y_2|y_1,i})^2 \quad (15)$$

where $\pi_{y_2|y_1,i} = P_i(Y_2 = 1|Y_1 = y_1)$ is the conditional probability that a loan is in default in the P2P lending platform given that it is $(y_1 = 1)$ or not $(y_1 = 0)$ reported in default by a credit bureau. n_1 is the number of the P2P defaults. The symbol + in the equation (15) indicates that the MSE is computed only for P2P defaulted loans, coherently with Andreeva et al. (2016). From the results of the study on real data shown in Appendix B we obtain that the same values of (τ_1, τ_2) minimise the MSE_+ for different dependence structure (Gaussian, Clayton, Gumbel, Frank and Joe copulas).

We choose a copula function among the alternatives defined in Table 1 to define a given dependence structure in the model. To choose the best copula among the provided competitors, as suggested by Zimmer and Trivedi (2006) and Radice et al. (2016), we apply either the Akaike information Criterion (AIC) or Schwarz Bayesian Information Criterion (BIC) on the training set. These criteria are based on the log-likelihood function of the model defined in equation (14) (see Breymann et al., 2003). In our context:

$$AIC = 2k - 2l(\delta)$$

$$BIC = \ln(n)k - 2l(\delta)$$

where $l(\delta)$ is the maximised log-likelihood function, k is the number of estimated parameters, and n is the sample size. According to these criteria, the best fitting model is the one that minimises AIC **or** BIC. **Note that when all rival models have the same number of parameters, it is equivalent to use the log-likelihood, AIC or BIC for the model selection (see Panagiotelis et al., 2017)**.

To obtain stable computations, the Fisher scoring is advisable for simultaneous equation estimation methods (Marra and Radice, 2017). Particularly, we implement the Fisher scoring in the `BivGEV` R package using a trust region method based on the Fisher matrix.

To derive the Fisher information matrix, we consider the following derivatives:

$$\begin{aligned}
\frac{\vartheta \pi_{11}}{\vartheta \boldsymbol{\beta}_j} &= \frac{\vartheta C_\lambda}{\vartheta \pi_j}\frac{\vartheta \pi_j}{\vartheta \boldsymbol{\beta}_j} \quad j = 1, 2 \\
\frac{\vartheta \pi_{10}}{\vartheta \boldsymbol{\beta}_1} &= \left(1 - \frac{\vartheta C_\lambda}{\vartheta \pi_1}\right)\frac{\vartheta \pi_1}{\vartheta \boldsymbol{\beta}_1} \\
\frac{\vartheta \pi_{10}}{\vartheta \boldsymbol{\beta}_2} &= -\frac{\vartheta C_\lambda}{\vartheta \pi_2}\frac{\vartheta \pi_2}{\vartheta \boldsymbol{\beta}_2} \\
\frac{\vartheta \pi_{01}}{\vartheta \boldsymbol{\beta}_2} &= \left(1 - \frac{\vartheta C_\lambda}{\vartheta \pi_2}\right)\frac{\vartheta \pi_2}{\vartheta \boldsymbol{\beta}_2} \\
\frac{\vartheta \pi_{01}}{\vartheta \boldsymbol{\beta}_1} &= -\frac{\vartheta C_\lambda}{\vartheta \pi_1}\frac{\vartheta \pi_1}{\vartheta \boldsymbol{\beta}_1} \\
\frac{\vartheta \pi_{11}}{\vartheta \lambda} &= -\frac{\vartheta \pi_{10}}{\vartheta \lambda} = -\frac{\vartheta \pi_{01}}{\vartheta \lambda} = \frac{\vartheta C_\lambda}{\vartheta \lambda}
\end{aligned}$$

The Fisher information matrix is

$$-E\left(\frac{\vartheta^2 l}{\vartheta \boldsymbol{\delta} \vartheta \boldsymbol{\delta}^T}\right) = \frac{\vartheta p}{\vartheta \boldsymbol{\delta}}\left(diag(p) + \frac{1}{1 - \pi_{11} - \pi_{10} - \pi_{01}}\mathbf{1}\mathbf{1}^T\right)\frac{\vartheta p}{\vartheta}\boldsymbol{\delta}^T$$

with $\boldsymbol{\delta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \lambda)$, $p = (\pi_{11}, \pi_{10}, \pi_{01})$ and $\mathbf{1} = (1, 1, 1)^T$.

3. Simulation study

We perform a Monte Carlo simulation study to explore the empirical properties of the BivGEV model. All computations are carried out using the `BivGEV` package in the `R` environment available in GitHub.

The simulation study is based on the following bivariate model

$$y_{1i}^* = \alpha_1 + x_{1i}\beta_1 + \epsilon_{1i} \tag{16}$$
$$y_{2i}^* = \alpha_2 + x_{2i}\beta_2 + \epsilon_{2i} \tag{17}$$

where the binary outcomes y_{1i} and y_{2i} are obtained based on the equations (13).

We generate the values for the covariate x_1 and x_2 from a normal distribution $N(1.2, 0.25)$. The parameters of interest are set as follows: in the equation (16) α_1=-1.7 and β_1=0.4, while in the equation (17) α_2=-2.2 and β_2=0.8. We choose these parameter values in order to obtain imbalanced proportions of 1s and 0s for the dependent variables Y_1 and Y_2 (around 6% of ones in both the equations). Following the results of Theorem (2.1), we generate the error terms ϵ_1 and ϵ_2 in the equations (16) and (17) from the two cumulative distribution functions $F_{\epsilon_j}(\varepsilon) = 1 - H(-\varepsilon)$ defined in Theorem 2.1, where $H(\varepsilon) = exp\{1 - (1 + \tau\varepsilon)^{(-1/\tau)}\}$ with parameters $\tau_1 = \tau_2 = -0.3$.

We choose the Gaussian and the survival Gumbel copula† to generate the dependence structure between the error terms ϵ_1 and ϵ_2. We consider the values for the association parameter λ that correspond to a Kendall's Tau coefficient of 0.2, 0.5 and 0.8, respectively, as explained in Section 2.2. We use the function `BiCopSim(n, family, lambda)` implemented in the R package `VineCopula` (Nagler et al., 2017) to represent the dependence structure. We consider different sample sizes $n = 1,000; 3,000; 5,000$ and $1,000$ replications for each combination of parameter settings.

We analyse the distribution of the estimates of the parameters β_1 and β_2, defined in the equations (16) and (17), and the Kendall's Tau coefficient defined in equation (7). As Table 1 shows, the copula parameter λ is defined on different intervals for different copula function. To make the results comparable between different copulas, we instead analyse the estimates of the Kendall's Tau coefficient.

Figure 3 and Figure 4 present the boxplots of the estimates of the parameters β_1, β_2 and of the Kendall's Tau coefficient for three sample sizes $n = 1,000; 3,000$ and $5,000$. We show the results only for the Gaussian copula in Figure 3 and for the Gumbel copula in Figure 4 as they are coherent with those obtained for other copula families‡.

For both the copula functions and for all the parameters, as the sample size n increases, the estimates converge to their true values represented by the horizontal line in Figure 3 and Figure 4. In addition, the standard deviations of the parameter estimators become smaller for larger n. As the Kendall's Tau coefficient increases, the precision of the estimates improves. This means that β_1 and β_2 are more accurately estimated as the association between the dependent variables Y_1 and Y_2 increases. Even if this result

†From the Theorem (2.1) we obtain that a survival Gumbel copula between the error terms is equivalent to a Gumbel copula between the marginal default probabilities.

‡The simulation results for other copula functions are available upon request from the authors.

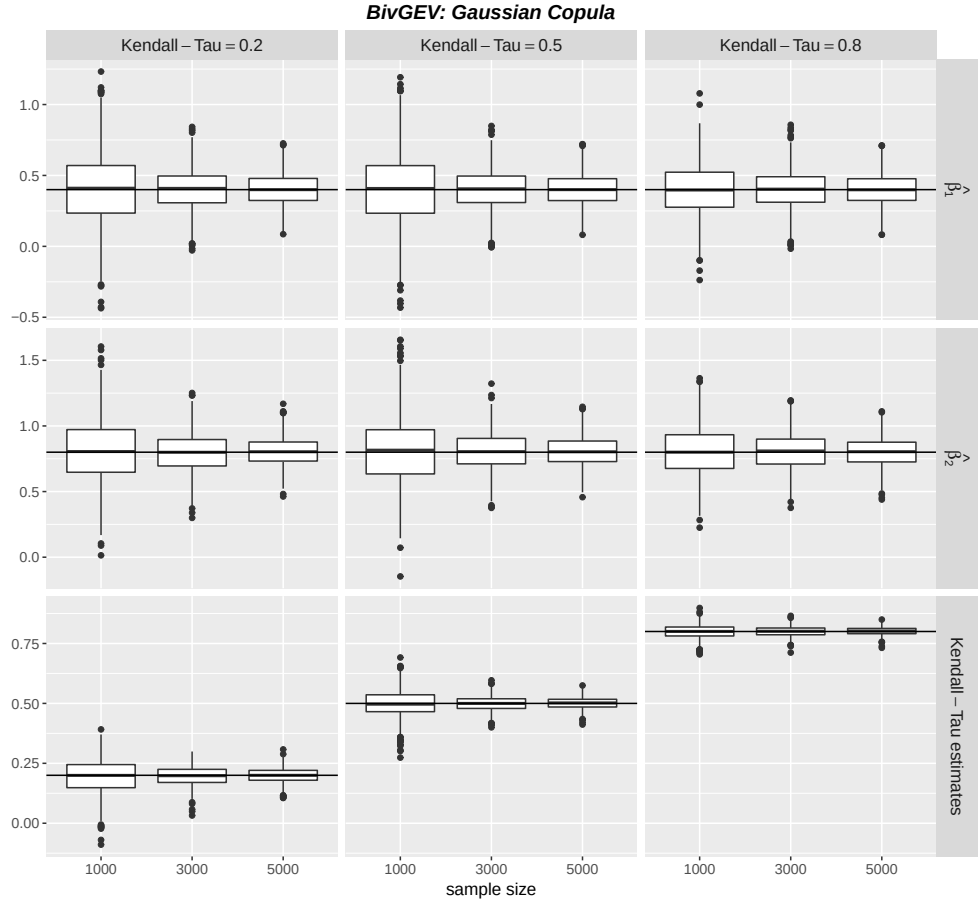

Figure 3. Boxplots of the estimates of β_1, β_2 and of the Kendall's Tau coefficient. Data are generated from a Gaussian copula with different values of Kendall's Tau coefficient = $0.2, 0.5, 0.8$ and different sample sizes $n = 1,000; 3,000; 5,000$. The number of Monte Carlo replications is $1,000$.

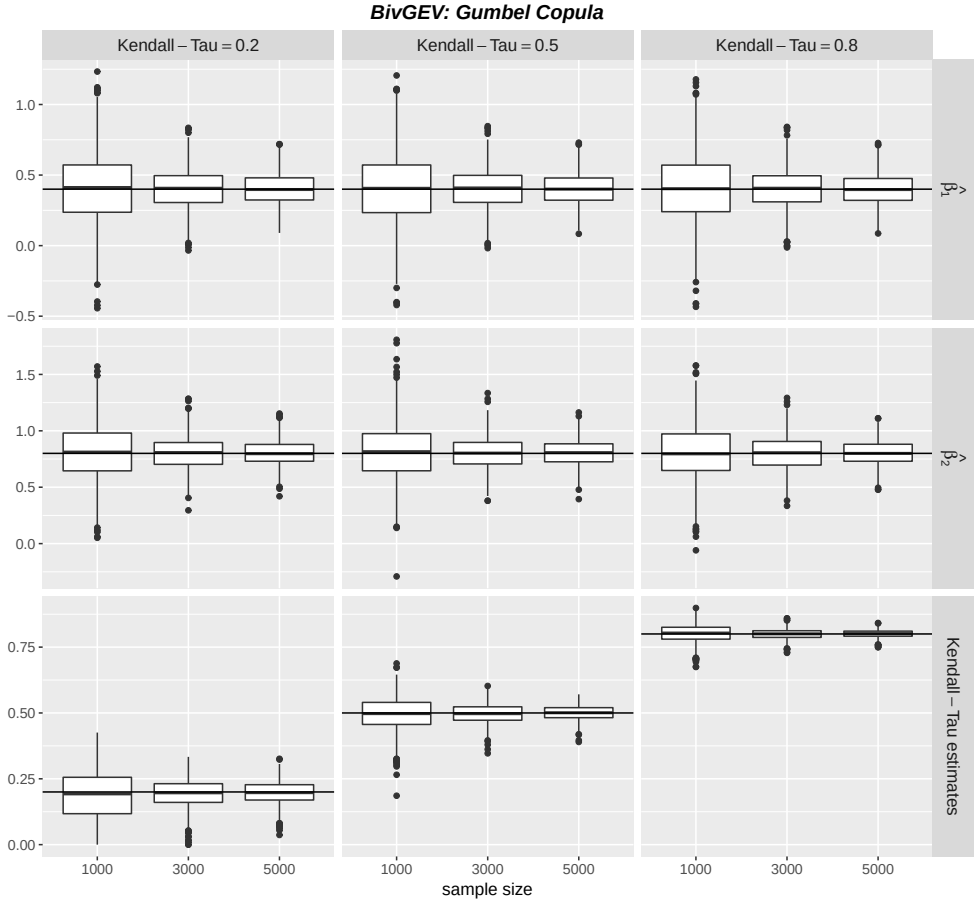

Figure 4. Boxplots of the estimates of β_1, β_2 and of the Kendall's Tau coefficient. Data are generated from a Gumbel copula with different values of Kendall's Tau coefficient = $0.2, 0.5, 0.8$ and different sample sizes $n = 1,000; 3,000; 5,000$. The number of Monte Carlo replications is $1,000$.

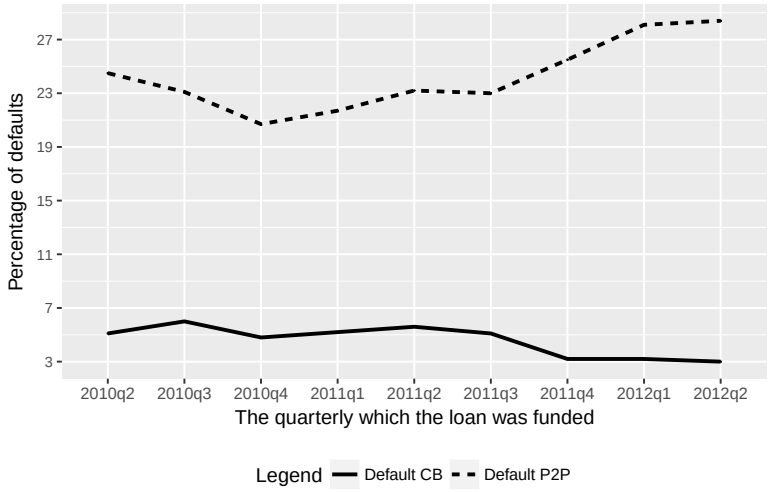

Figure 5. Quarterly time series of the percentage of defaults for the credit bureau (CB) and for P2P loans provided by Lending Club.

might appear counter-intuitive, a lower level of association between Y_1 and Y_2 suggests that the precision in the parameter estimates is lower. In other terms, if the association between the dependent variables Y_1 and Y_2 is high, then this can be easily detected by a bivariate model that estimates jointly all the parameters. Similar results have been obtained by Chib and Greenberg (2007), Marra and Radice (2011).

4. Empirical analysis

4.1. Data

To improve lender screening and monitoring (Miller, 2015), some P2P lending platforms make the historical information of the loans that have been funded publicly available, jointly with the characteristics of the loans and their status (default or non-default). For this analysis, we use data provided by the biggest US P2P lending platform, Lending Club§, who issued in total $29 billion in loans as of the second quarter of 2017¶. At the beginning, we consider all the loans funded by Lending Club. We choose only 60 months loans to have a time horizon that is sufficiently long.

As data starts from 2010, we remove all loans that are still outstanding. This leaves a sample of 18,113 loans, from the second quarter of 2010 to the second quarter of 2012. We report in Figure 5 the time series of the default rates for the P2P platform and for the credit bureaus for each quarter.

To better understand the relationship between being in default for a credit bureau and for a P2P lending platform, we propose to use a bivariate regression model whose dependent variables are:

§The data are available in https://www.lendingclub.com/info/download-data.action
¶https://www.lendingclub.com/info/statistics.action

- Credit bureau default (Y_1): a binary variable that takes the value of 1 if the borrower has one or more public record bankruptcies and 0 otherwise.
- P2P default (Y_2): a binary variable that takes the value of 1 if the borrower at the end of the contractual terms is in the status of default or charged-off in the Lending Club platform, and 0 otherwise.

Given the information provided by Lending Club, the predictors of the scoring model for the credit bureau default and for the P2P default are:

- *Loan purpose*: We classify the loan purposes as follows: debt consolidation; credit card; car financing; home improvement; major purchase (including others); personal (including medical, moving, vacation, wedding, educational) and small business. We use 6 dummy variables to represent these options.
- *Housing situation*: The home ownership status is information provided by the borrower during registration or that it is obtained from the credit report. A borrower is classified as having a mortgage, being a renter, owning their home/other situation using two dummy variables.
- *Interest rate*: Interest Rate on the loan funded.
- *Annual income*: The self-reported annual income in US dollars provided by the borrower.
- *Revolving utilisation*: measures the revolving line utilisation rate.
- *DTI*: The ratio of a borrower's total monthly debt payments (excluding mortgage and the requested Lending Club loan) to a borrower's self-reported monthly income.
- *Credit history length*: Number of years since a borrower first opened a credit line.
- *Loan amount to annual income*: The ratio between the loan amount and the annual income.
- *Spatial variables defined using the first digit of the ZIP Code*: This feature is represented using nine binary variables. ‖

To avoid multicollinearity, we consider only the predictor variables with a Variance Inflation Factor (VIF) lower than 5. As the VIF is about the explanatory variables, it can be used also for binary outcomes, as previous publication show (e.g. Allison, 2012: pages 60-63; Calabrese and Osmetti, 2013; Calabrese et al., 2015, Andreeva et al., 2016).

4.2. Estimation results

In this section we present the main results from the application of the BivGEV model to credit scoring in P2P lending market. This includes the model selection and estimation, the interpretation of the parameter estimates and the comparison of the predictive accuracy with those of other models traditionally used in the industry and in academic

‖The adopted classification is reported in Appendix A.

Table 2. Copula parameter estimates with asymptotic confidence intervals and the corresponding values of the Kendall's Tau coeff

Copula	Copula parameter λ	Confidence intervals	Kendall's Tau	Confidence intervals
Gaussian	0.133	$(0.087; 0.174)$	0.085	$(0.056; 0.111)$
Clayton	0.091	$(0.055; 0.147)$	0.044	$(0.027; 0.069)$
Gumbel	1.140	$(1.100; 1.200)$	0.122	$(0.088; 0.169)$
Frank	0.983	$(0.530; 1.410)$	0.108	$(0.059; 0.154)$
Joe	1.450	$(1.270; 1.730)$	0.201	$(0.132; 0.289)$

research. We fit the BivGEV model proposed in Section 2.3 on the data from Lending Club described in Section 4.1. The sample size is 18,113. The percentage of defaulted P2P loans is 25.18% and the percentage of credit bureau defaults is 4.17%.

To estimate the BivGEV model we apply the procedure described in Section 2.4. First, we split the sample in a training sample (80% of the observations) and a control sample (20%). Then, we choose the values for the parameters τ_1 and τ_2 of the GEV marginal models. As explained in Section 2.1, when the sample is imbalanced with low percentage of defaults ($y = 1$), the best link function to predict defaults is obtained for $\tau < 0$ (Weibull class). To check this evidence in our context, we estimate the BivGEV model for different values of τ_1 and τ_2 in the range $[-0.85; 0.35]$ and different copula functions (Gaussian, Gumbel, Clayton, Frank and Joe). **We choose the range** $[-0.85; 0.35]$ **for the parameters** τ_1 **and** τ_2 **because Tables 8-12 in Appendix B show that the performances of the BivGEV model are worse for values of** τ_1 **and** τ_2 **close to** -0.85 **and** 0.35**. We estimate all the BivGEV models by using the R package *BivGEV* publicly available in GitHub.**

As described in Section 2.4, for each copula function, we select the values of τ_1 and τ_2 that minimise the MSE_+ computed on the training sample. In this way we select the model with the best predictive accuracy in predicting defaults (e.g., Andreeva et al., 2016, Calabrese et al., 2016, and Calabrese and Osmetti, 2013). Tables 8-12 in Appendix B report an extract of the MSE_+ obtained for different values of τ_1 and τ_2 and for each copula function††. We observe that the model with the best predictive accuracy is obtained for $\tau_1 = -0.75$ and $\tau_2 = -0.15$ for each copula function.

Chosen the values of $\tau_1 = -0.75$ and $\tau_2 = -0.15$, we estimate the parameter λ of the copula functions **by maximising the equation (14)**. We also compute the Kendall's Tau coefficient using the equation (7) and the asymptotic confidence intervals of these parameters. We report the results in Table 2.

The confidence intervals for the copula parameters and for the Kendall's Tau coefficient are obtained by Bayesian posterior simulation following the procedure described in Radice et al. (2016) pag. 988.

The estimate of the parameter λ for the Gaussian copula is close to zero (0.133), describing low dependence (linear correlation) between the marginal probabilities. This result can be due to the type of dependence structure between the marginal probabilities in a Gaussian copula that can be only linear and not in the tails. To analyse tail dependence, we consider the Clayton, the Gumbel and the Joe copulas. In Table 2, we

$**https://github.com/BivGEV/BivGEV$. In the supplementary material, we provide the R code to guide a researcher or practitioner in the estimation of a BivGEV model.

††The other results are available upon request.

Table 3. Model selection measures of the BivGEV models for different copula families. In bold the selected model

Copula	*AIC*	*BIC*
Gaussian	20,229.15	20,441.43
Clayton	20,230.27	20,442.54
Gumbel	**20,226.41**	**20,438.68**
Frank	20,226.47	20,438.71
Joe	20,226.57	20,438.85

obtain that the copula parameter λ is 0.091 and the Kendall's Tau coefficient is 0.044 for the Clayton copula. The latter value indicates a low level of association between the marginal default probabilities. We compute the lower tail dependence parameter for the Clayton copula using equation (8) and we obtain a value close to zero ($\chi_u \simeq 0$). Therefore, the dependence between lower values of default probabilities for the P2P platform and for credit bureaus is negligible. Both the Gumbel and the Joe copulas show upper tail dependence between the marginal distributions. The Kendall's Tau coefficient for the Gumbel copula is lower (0.122) than that of the Joe copula (0.201). The upper tail dependence parameters for the Gumbel copula (χ_u=0.163) and for the Joe copula ($\chi_u = 0.387$) show, respectively, a low and a medium intensity of association between the two default probabilities.

To choose the copula with the best fit among the provided alternatives, we consider two main criteria. As the copula models are non-nested, we use the Akaike Information Criterion and the Bayesian information criterion (see Section 2.4). We report the results in Table 4.2.

We find that the copula that best fits the data according to both the AIC and the BIC is the Gumbel copula. This result is in line with the expectations of obtaining an upper tail dependence between the two default probabilities, coherently with the substantial empirical evidence for interaction between default events. Dependence between defaults stems from different sources. Common macroeconomic factors such changes in economic growth can affect different borrowers. This type of dependence between defaults has been modelled both in reduced models (Duffie and Singleton, 2003; Lando, 2004) and structural models (Vasicek, 2002) and included in the Basel II Accord (BCBS, 2006). Moreover, dependence between defaults could be caused by direct economic links between borrowers. These direct links lead to default contagion and counterparty risk, which has generated a lot of interest in the recent literature (Calabrese et al., 2017).

We report in Table 4 the parameter estimates of the BivGEV model for the Gumbel copula applied to the training sample of 14,490 loans. We compare them with those obtained from an univariate logit model where the dependent variable is default or non-default in the P2P lending platform. **In the first Logit model (Logit$_1$ model) we use the same explanatory variables of the BivGEV model. In the second Logit model (Logit$_2$ model), we add whether the loan has been classified in default or not by a credit bureau as an additional explanatory variable. We consider the Logit$_1$ and Logit$_2$ models because they are commonly used in retail credit risk management (Baesens et al., 2016; Lin et al., 2013; Thomas et al., 2017). Initially, we include all the independent variables and then**

Table 4. Parameter estimates for the BivGEV model with Gumbel copula and for the univariate Logit_2 model. $^{*}p - value \leq 0.1$ $^{**}p - value \leq 0.05$

	BivGEV model						$Logit_2$ model		
	Eq. 1: credit bureau default			Eq. 2: P2P default			Eq. P2P default		
Variables	Estimate	Std. Error	P-value	Estimate	Std. Error	P-value	Estimate	Std. Error	P-value
Loan Purpose									
Credit card				-0.007^{*}	0.037	0.059	-0.131^{**}	0.067	0.000
Car financing				-0.243^{**}	0.062	0.000	-0.512^{**}	0.130	0.000
House				-0.077^{*}	0.040	0.084	-0.146^{*}	0.075	0.051
Major purchase	-0.387^{**}	0.108	0.000						
Small business				0.421^{**}	0.050	0.000	0.705^{**}	0.081	0.000
Housing situation									
Rent	-0.282^{**}	0.067	0.000						
Own				0.083^{*}	0.043	0.057	0.136^{*}	0.075	0.071
Borrower assessment									
Interest rate	0.110^{**}	0.009	0.000	0.058^{**}	0.004	0.000	0.099^{**}	0.006	0.000
Borrower characteristics									
ln(Annual income)	-0.655^{**}	0.065	0.000	-0.327^{**}	0.027	0.000	-0.559^{**}	0.049	0.000
Credit history									
Revolving utilisation				0.002^{**}	0.001	0.000	0.003^{**}	0.001	0.000
Credit history length	0.024^{**}	0.005	0.000	0.007^{**}	0.002	0.000	0.011^{**}	0.003	0.001
Borrower indebtedness									
Loan amount to annual income	-2.307^{**}	0.280	0.000	0.286^{**}	0.099	0.004	0.602^{**}	0.173	0.000
Spatial variables									
Zone 0	-0.269^{**}	0.115	0.019						
Zone 1									
Zone 2				-0.061^{*}	0.036	0.009	-0.108^{**}	0.064	0.009
Zone 3									
Zone 4	0.327^{**}	0.099	0.001						
Zone 5									
Zone 6				-0.139^{**}	0.046	0.000	-0.245^{**}	0.084	0.003
Zone 7				-0.067^{*}	0.038	0.007	-0.137^{*}	0.070	0.051
Zone 8									
Zone 9									
Credit bureau default							0.467^{**}	0.089	0.000
Intercept	3.780^{**}	0.739	0.000	2.115^{**}	0.297	0.000	3.037^{**}	0.526	0.000
Link function	GEV			GEV			Logit		
τ_1	-0.75								
τ_2				-0.15					
Number of observations	14,490						14,490		

we remove those that are not statistically significant at a confidence level $\alpha = 0.10$.

Table 4 shows that the determinants of the default in P2P lending and for credit bureau are different. For example, most of the spatial dummy variables are significant for P2P default but not for default reported by credit bureaus. For housing, being a renter is an important predictor for the credit bureau default. On the other hand, the probability of P2P default is lower for home owners, in line with expectations and with results obtained by Serrano-Cinca et al. (2015) on data provided by Lending Club.

It is interesting that the ratio between loan amount and annual income has a negative relationship with credit bureau default and a positive one with P2P default. From the first result, we could deduce that borrowers that are reported in default by a credit bureau usually apply for lower loan amounts to annual income than in the P2P platform. However, higher loan amounts to income in the P2P platform show higher default probabilities, in agreement with Serrano-Cinca et al. (2015). The parameter estimates for P2P default obtained using the BivGEV and the univariate Logit_2 model are similar. As expected, in the second model being in credit bureau default is a highly signifi-

cant explanatory variable that is positively associated to the probability of P2P default. Analogously to most of the studies on P2P lending (e.g. Dorfleitner et al., 2016; Freedman and Jin, 2008; Serrano-Cinca et al., 2015), we find that the interest rate being charged is a highly significant predictor of P2P default. Also, annual income and revolving utilisation are important for explaining P2P default, as Emekter et al. (2015) and Serrano-Cinca et al. (2015) have previously shown. If we focus our attention on the loan purpose, we obtain interesting results in accordance with Serrano-Cinca et al. (2015): loans for small businesses show a positive relationship with P2P default. This is negative for credit card, car financing and house loans.

4.3. Predictive accuracy

In this section, we compare the BivGEV model with the bivariate probit proposed by Winkelmann (2011) using the Gumbel copula with two univariate logit models one includes the dummy variable credit bureau default **as explanatory variable ($\mathbf{Logit}_2$) and the other one no ($\mathbf{Logit}_1$)**. Moreover, since we are working with imbalanced sample, we apply the SMOTE approach to Logit_1 and Logit_2 models. The SMOTE (Chawla et al., 2002; Baesens et al., 2016) is a synthetic minority over-sampling technique that deals with the imbalanced sample by producing an adjusted dependent variable to be used in the credit scoring model‡‡. In order to consider well-calibrated probabilities of default compliant to Basel III (Basel Committee on Banking Supervision, 2015), we apply the adjust of the posterior default probabilities proposed in Saerens et al. (2002) and Baesens et al. (2016).

To fit the bivariate probit we use the R package *GJRM* (Marra and Radice, 2017), while to apply the SMOTE technique we use the R package *DMwR* (Torgo, 2013).

As previously mentioned, to avoid overfitting we compute the measures of predictive accuracy on an out-of-sample of 20% of observations randomly drawn from the data set. First, we estimate the conditional probability $\pi_{y_2|y_1,i} = P_i(Y_2 = 1|Y_1 = y_1)$ that a loan is in default in the P2P lending platform, given that it is reported in default ($y_1 = 1$) or not ($y_1 = 0$) by a credit bureau. For assessing the predictive accuracy, we compute the MSE_+, the Area Under the Curve (AUC), the H-measure with a severity ratio of 0.01 (Hand, 2009 and 2010) using $\pi_{y_2|y_1,i}$ for the bivariate models and $\pi_{y_{2i}}$ for the univariate models. We report the values of MSE_+, the AUC and the H-measure in Table 5.

The lower the values of the MSE_+ or the higher the values of the AUC and the H-measure, the more accurate the model correctly classifying defaults and non-defaults in the P2P lending platform.

Firstly, we compare the accuracy measures of **$\mathbf{Logit}_1$, $\mathbf{Logit}_2$**, BivProbit and BivGev models. As expected, the model with the worst performance is the univariate **Logit** model without credit bureau default as an explanatory variable **($\mathbf{Logit}_1$)**. Most of the performance measures in Table 5 show that the approach suggested in this paper of using a BivGEV model outperforms the univariate **$\mathbf{Logit}_1$** commonly used in industry. Coherently with the results obtained in the univariate framework (Andreeva et al., 2016;

‡‡For lack of space we did not report the parameter estimates obtained using the SMOTE technique. They are available upon authors' request. However, we highlight that some of the estimates are incoherent with those obtained in Table 4 and with the expectations

Table 5. Predictive accuracy measures for different models. In bold the selected model

Model	MSE_+	AUC	$H_{0.01}$
Logit_1	0.51399	0.64641	0.00265
Logit_2	0.51254	0.64836	0.00264
SMOTE + Logit_1	0.52381	0.63280	0.00428
SMOTE + Logit_2	0.52200	0.63396	0.00427
$\text{BivProbit}_{(Gumbel)}$	0.51220	0.64886	0.00306
$\text{BivGEV}_{(Gumbel)}$	**0.51196**	**0.64902**	**0.00495**

Calabrese et al., 2016; Calabrese and Osmetti, 2013 and 2015), using an asymmetric link function improves the predictive accuracy of the scoring model, as can be seen by comparing the results of the BivGEV and the bivariate probit model. Finally, we observe that the BivGev model is more accurate in forecasting defaults than the two **Logit models (SMOTE+Logit_1 and SMOTE+Logit_2)** when a SMOTE approach is applied.

5. Estimating the loss for the P2P lending platform

In this section, we assess the impact of using different scoring models for P2P loans on the Value at Risk (VaR). The VaR at the confidence level $(1-\alpha)$ is the level of losses on the portfolio that will be exceeded $(1-\alpha)\%$ of time on average (Thomas et al., 2017). Compliant to Basel III and IFRS 9, we compute the Expected Credit Loss as

$$EL = PD \times LGD \times EAD \tag{18}$$

where

- PD is the probability of default estimated using a regression model analysed in section 3.3;
- EAD is the Exposure At Default;
- LGD is the Loss Given Default computed as 1−(amount recovered - recovery fee)/EAD.

To compute the EAD and the LGD we use the Lending Club data. As the LGD is zero for non-defaulted P2P loans, the EL for them is also zero. For this reason, we compute the EL only on P2P defaults using the PDs obtained from **Logit_1, Logit_2, SMOTE+Logit_1, SMOTE+Logit_2**, BivProbit and BivGEV models with a Gumbel copula. The VaR at different confidence levels are reported in Table 6.

The confidence levels from 95% to 99.9% correspond to the usual intervals considered by banks. The main result is that, for all the confidence levels except 95%, the VaR estimates of the various models satisfy the same ordering: SMOTE + Logit_1 always underestimates the credit VaR of Logit_2 which, in turn, underestimates the VaR obtained Logit1 which underestimates the VaR obtained using a bivariate model. Particularly, the BivGEV model proposed always shows the highest VaR, excluding 95%.

Given the disadvantages of the VaR (Saunders and Allen, 2010), we also compute the Expected Shortfall (Acerbi and Tasche, 2002). The Expected Shortfall (ES) at the

Table 6. Value at Risk of the (expected credit) loss distribution for different confidence levels

Model	*95th percentile*	*97th percentile*	*99th percentile*	*99.9th percentile*
Logit$_1$	**8,843**	**9,681**	**12,551**	**15,726**
Logit$_2$	**8,935**	**9,659**	**12,538**	**15,618**
SMOTE + Logit$_1$	**9,042**	**9,563**	**12,358**	**15,237**
SMOTE + Logit$_2$	**9,084**	**9,689**	**12,580**	**15,990**
BivProbit$_{(Gumbel)}$	8,944	9,738	12,618	16,120
BivGEV$_{(Gumbel)}$	8,966	9,756	12,714	16,320

Table 7. Expected Shortfall of the (expected credit) loss distribution for different confidence levels

Model	*95th percentile*	*97th percentile*	*99th percentile*	*99.9th percentile*
Logit$_1$	**10,849**	**11,880**	**14,493**	**16,229**
Logit$_2$	**10,963**	**12,032**	**14,669**	**16,698**
SMOTE + Logit$_1$	**10,958**	**11,995**	**14,108**	**16,620**
SMOTE + Logit$_2$	**10,997**	**12,066**	**14,253**	**16,558**
BivProbit$_{(Gumbel)}$	10,990	12,069	14,698	17,021
BivGEV$_{(Gumbel)}$	11,054	12,147	14,814	17,125

confidence level $(1-\alpha)$ is the average of all losses which are greater or equal than the VaR at the confidence level $(1-\alpha)$. The results reported in Table 7 show that the ES estimates of the bivariate models are higher than the estimates of the univariate models for all the confidence levels except 95%. In particular, the BivGEV model always shows the highest ES. As the LGD and EAD in equation (18) show the same values for the six models in Table 6 and 7, we can deduce that the BivGEV model overcomes the drawback of the **univariate logit** and bivariate probit models in underestimating the probability of default for the actual defaults.

6. Concluding remarks

We suggested a scoring model for P2P lending. Firstly, we proposed to use a bivariate approach that jointly modelled the defaults on the P2P platform and credit bureau defaults. Secondly, we introduced the BivGEV model based on an asymmetric link function defined using the quantile function of a GEV random variable. **The model is implemented in the R package `BivGEV` available in GitHub. In the supplementary material of this paper, we report the `R` code that can be useful for researchers and practitioners who are interested in fitting the BivGEV model. Monte Carlo experiments were conducted to evaluate the asymptotic properties of the BivGEV estimates.** We applied our proposal and its competitors to data on 18,113 P2P loans provided by Lending Club from 2010 to 2012. The main advantage of the BivGEV model lies in its capacity to better forecasting P2P defaulted loans. In the empirical analysis we showed that the BivGEV model provides more accurate estimates of the VaR and the Expected Shortfall. This means that P2P platforms, based on our proposal, could improve their internal assessment. Future research might look at the possibility of applying a Bayesian approach to estimate all the parameters of the BivGEV model. Another possible extension could be to use a regression splines in the BivGEV model to flexibly estimate the covariate effects (Radice et al. 2016). **Finally,**

it would be interesting to apply the BivGEV model on two simultaneous definitions of default, instead of having one definition that is sequential to the other one as it occurs in Lending Club data.

References

Acerbi C., Tasche D. (2002) Expected Shortfall: a natural coherent alternative to Value at Risk. *Economic Notes*, 31, 379–388.

Agresti A. (2002) *Categorical Data Analysis*, Wiley, New York.

Allison P. D. (2012) *Logistic Regression Using SAS: Theory and Application*, SAS Institute Inc.

Andreeva G., Calabrese R., Osmetti S.A. (2016) A comparative analysis of the UK and Italian small businesses using Generalised Extreme Value models. *European Journal of Operational Research*, 249(2), 506–516.

Baesens B., Rösch D., Scheule H. (2016) *Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS*, John Wiley & Sons Inc., New Jersey.

Bank of England (2018) *Credit Conditions Survey*, Survey results, First quarter.

Basel Committee on Banking Supervision (BCBS) (2006) *International convergence of capital measurement and capital standards: A revised framework*. Bank for International Settlements, Basel, June.

Basel Committee on Banking Supervision (BCBS) (2015) *Guidelines on accounting for expected credit losses. Retrievedfromhttp* : *//www.bis.org/bcbs/publ/d*311*.pdf*.

Breymann W., Dias A., Embrechts P. (2003) Dependence structures for multivariate high-frequency data in finance. *Quantitative Finance*, 3, 1–14.

Cai S., Lin X., Xu D., Fu X. (2016) Judging online peer-to-peer lending behavior: A comparison of first-time and repeated borrowing requests. *Information & Management*, 53, 857–867.

Calabrese R., Degl'Innocenti M., Osmetti S.A. (2017) The effectiveness of TARP-CPP on the US banking industry. A new copula-based approach. *European Journal of Operational Research*, 256(3), 1029–1037.

Calabrese R., Marra G., Osmetti S. A. (2016) Bankruptcy prediction of small and medium enterprises using a flexible binary generalized extreme value model. *Journal of the Operational Research Society*, 67(4), 604–615.

Calabrese R., Giudici P. (2015) Estimating bank default with generalised extreme value models. *Journal of the Operational Research Society*, 66, 1783–1792.

Calabrese R., Osmetti S.A. (2015) Improving Forecast of Binary Rare Events: A GAM-Based Approach. *Journal of Forecasting*, 34(3), 230–239.

Calabrese R., Osmetti S.A. (2013) Modelling small and medium enterprise loan defaults as rare events: the generalized extreme value regression model. *Journal of Applied Statistics*, 40(6), 1172–1188.

Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeywer W.P. (2002) SMOTE: Synthetic minority over-sampling techniques. *Journal of Artificial Intelligence Research*, 16, 321–357.

Chib S., Greenberg E. (2007) Semiparametric modeling and estimation of instrumental variable models. *Journal of Computational and Graphical Statistics*, 16, 86–114.

ECB (2018) *Euro area bank lending survey*. Second quarter 2018.

Cole R. A. (2012) How Did the Financial Crisis Affect Small Business Lending in the United States?. *Working Paper, Small Business Administration*, SBAHQ-10-M-0208.

Duffie D., Kenneth J.S. (2003) *Credit Risk Pricing, Measurement and Management*, Princeton University Press.

Dorfleitner G., Priberny C., Schuster S., Stoiber J., Weber M., de Castro I., Kammler J. (2016) Description-text related soft information in peer-to-peer lending – Evidence from two leading European platforms. *Journal of Banking and Finance*, 64, 169–187.

Emekter R., Tu Y., Jirasakuldech B., Lu M. (2015) Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. *Applied Economics*, 47(1), 54–57.

Falk M. , Hüsler J., Reiss R.D. (2010) *Laws of small numbers: Extremes and rare events*, 3rd ed., Springer, Basel.

Fermanian J.D. (2005) Goodness-of-fit tests for copulas. *Journal of Multivariate Analysis*, 95(1), 119–152.

Fisher N.I (1997) *Copulas*, Encyclopedia of Statistical Sciences, 159-163. ed. S. Kotz, C.B. Read, D.L. Banks, Wiley, New York.

Freedman S., Jin G.Z (2008) Do social networks solve information problems for peer-to-peer lending? Evidence from prosper.com. *SSRN Working Paper, NET Institute*, Working Paper No. 08–43

Genest C., Nikoloulopoulos A.K., Rivest L.-P., Fortin M. (2013) Predicting dependent binary outcomes through logistic regressions and meta-elliptical copulas. *Brazilian Journal of Probability and Statistics*, 27, 265–284.

Greene, W.H. (2012) *Econometric Analysis*, Prentice Hall, New York.

Guo, Y., Zhou, W., Luo, C., Liu, C., Xiong, H. (2016) Instance-based credit risk assessment for investment decisions in P2P lending. *European Journal of Operational Research*, 249(2), 417–426.

Hand D. J.(2009) Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, 77, 103–123.

Hand D.J. (2010) Evaluating diagnostic tests: the area under the ROC curve and the balance of errors. *Statistics in Medicine*, 29 , 1502–1510.

Joe H. (1997) *Multivariate Models and Dependence Concepts*, Chapman & Hall, Boca Raton.

King G., Zeng L. (2001) Logistic Regression in Rare Events Data. *Political Analysis*, 9, 137–163.

Kotz S., Nadarajah S. (2000) *Extreme value distributions. Theory and applications*, Imperial College Press, London.

Lando, D. (2004) *Credit risk modeling: Theory and applications*, Princeton University Press.

Lin M., Prabhala N. R., Viswanathan S. (2013) Judging Borrowers by the Company They Keep: Friendship Networks and Information Asymmetry in Online Peer-to-Peer Lending. *Management Science*, 59(1), 17–35.

Marra G., Radice R. (2011) Estimation of a semiparametric recursive bivariate probit model in the presence of endogeneity. *The Canadian Journal of Statistics*, 39(2), 259–279.

Marra G., Radice R. (2017) *Generalised Joint Regression Modelling*. R package.

Miller S. (2015) Information and default in consumer credit markets: Evidence from a natural experiment. *Journal of Financial Intermediation*, 24(1), 45–70.

Nelsen R.B. (2006) *An Introduction to Copulas*. Springer, New York.

Nagler T., Schepsmeier U., Stoeber J., Brechmann E.C., Graeler B., Erhardt T., Almeida C., Min A, Czado C., Hofmann M., Killiches M, Joe H., Vatter T. (2017). *VineCopula - Statistical Inference of Vine Copulas*, Version: 2.1.3, R Package.

Owzar K., Sen P.K. (2003) Copulas: concepts and novel applications. *Metron*, LXI(3), 323–353.

Panagiotelis A., Czado C., Joe H., Stöber J. (2017) Model selection for discrete regular vine copulas. *Computational Statistics & Data Analysis*, 106, 138–152

Radice R., Marra G., Wojtys M. (2016) Copula regression spline models for binary outcomes. *Statistics and Computing*, 26(5), 981–995.

Saerens M., Latinne P., Decaestecker C. (2002) Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural Computation*, 14, 21–41.

Saunders A., Allen L. (2010) *Credit risk measurement in and out of the financial crisis*. John Wiley & Sons, New York.

Serrano-Cinca C., Gutièrrez-Nieto B., Lòpez-Palacios L. (2015) Determinants of default in P2P lending. *PLoS ONE* 10(10), 1–22.

Sklar A.W. (1959) *Fonctions de répartition à n dimension et leurs marges*, 229–231. Publ. Inst. Statist. Univ. Paris.

Smith R. L. (1989) Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone (with discussion). *Statistical Science*, 4, 367–393.

Thomas L., Crook J., Edelman D. (2017) *Credit scoring and its applications*. Mathematics in Industry, Second Edition.

Torgo L. (2013) *DMwR: Functions and data for Data Mining with R*. R package.

Trivedi P.K., Zimmer D.M. (2007) Copula modeling: an introduction for practitioners. *Foundations and Trends® in Econometrics*, 1(1), 1–111.

Vasicek, O. (2002) The Distribution of Loan Portfolio Value. *Risk*, 15, .

Winkelmann R. (2011) Copula bivariate probit models: with an application to medical expenditures. *Health Economics*, 21(12), 1444-1455.

Zhao T., Jones-Evans D. (2016) SMEs, banks and the spatial differentiation of access to finance. *Journal of Economic Geography*, 17, 791-824.

7. Appendix A

The United States are classified as follows:
Area 0: Connecticut (CT), Massachusetts (MA), Maine (ME), New Hampshire (NH), New Jersey (NJ), New York (NY, Fishers Island only), Rhode Island (RI), Vermont (VT), Virgin Islands (VI);
Zone 1: Delaware (DE), New York (NY), Pennsylvania (PA);
Zone 2: District of Columbia (DC), Maryland (MD), North Carolina (NC), South Carolina (SC), Virginia (VA), West Virginia (WV);
Zone 3: Alabama (AL), Florida (FL), Georgia (GA), Mississippi (MS), Tennessee (TN);
Zone 4: Indiana (IN), Kentucky (KY), Michigan (MI), Ohio (OH);
Zone 5: Iowa (IA), Minnesota (MN), Montana (MT), North Dakota (ND), South Dakota (SD), Wisconsin (WI);
Zone 6: Illinois (IL), Kansas (KS), Missouri (MO), Nebraska (NE);
Zone 7: Arkansas (AR), Louisiana (LA), Oklahoma (OK), Texas (TX);
Zone 8: Arizona (AZ), Colorado (CO), Idaho (ID), New Mexico (NM), Nevada (NV), Utah (UT), Wyoming (WY);
Zone 9: Alaska (AK), California (CA), Hawaii (HI), Marshall Islands (MH), Oregon (OR), Palau (PW), Washington (WA).

Table 8. Predictive accuracy measures and model selection measures of the BivGEV models with Gaussian copula for d values of (τ_1, τ_2). In bold the selected model

Credit bureau default	Default P2P	Predictive Accuracy	Model selection measures	
τ_1	τ_2	MSE_+	AIC	BIC
-0.75	-0.85	0.50398	20,246.84	20,459.11
-0.75	-0.75	0.50346	20,242.60	20,454.88
-0.75	-0.65	0.50303	20,238.98	20,451.25
-0.75	-0.55	0.50269	20,235.93	20,448.20
-0.75	-0.45	0.50242	20,233.44	20,445.71
-0.75	-0.35	0.50224	20,231.49	20,443.76
-0.75	-0.25	0.50212	20,230.07	20,442.34
-0.75	**-0.15**	**0.50208**	**20,229.15**	**20,441.43**
-0.75	-0.05	0.50210	20,228.74	20,441.02
-0.75	$+0.05$	0.50219	20,228.83	20,441.10
-0.75	$+0.15$	0.50234	20,229.40	20,441.67
-0.75	$+0.25$	0.50256	20,230.45	20,442.72
-0.75	$+0.35$	0.50284	20,231.98	20,444.26

8. Appendix B

We conduct a study on the BivGEV model selection following the procedure defined in Section 2.4. We consider the training set and the control set described in Section 4.1. Tables 8-12 report some results of the MSE_+ and the AIC and BIC calculated on the training sample. We consider several BivGEV models with 5 different copula functions (Gaussian, Gumbel, Clayton , Frank, Joe) and different combination of the values of τ_1 and τ_2 in the range $[-0.85; 0.35]$. The results are obtained by using the R package `BivGEV`.

Table 9. Predictive accuracy measures and model selection measures of the BivGEV models with Clayton copula for different values of (τ_1, τ_2). In bold the selected model

Credit bureau default	Default P2P	Predictive Accuracy	Model selection measures	
τ_1	τ_2	MSE_+	AIC	BIC
-0.75	-0.85	0.50411	20,248.06	20,460.33
-0.75	-0.75	0.50359	20,243.81	20,456.09
-0.75	-0.65	0.50316	20,240.17	20,452.45
-0.75	-0.55	0.50282	20,237.11	20,449.38
-0.75	-0.45	0.50255	20,234.61	20,446.88
-0.75	-0.35	0.50237	20,232.64	20,444.91
-0.75	-0.25	0.50225	20,231.20	20,443.47
-0.75	**-0.15**	**0.50220**	**20,230.27**	**20,442.54**
-0.75	-0.05	0.50223	20,229.84	20,442.11
-0.75	$+0.05$	0.50231	20,229.90	20,442.17
-0.75	$+0.15$	0.50247	20,230.44	20,442.72
-0.75	$+0.25$	0.50268	20,231.47	20,443.75
-0.75	$+0.35$	0.50296	20,232.98	20,445.25

Table 10. Predictive accuracy measures and model selection measures of the BivGEV models with Gumbel copula for different values of (τ_1, τ_2). In bold the selected model

Credit bureau default	Default P2P	Predictive Accuracy	Model selection measures	
τ_1	τ_2	MSE_+	AIC	BIC
-0.75	-0.85	0.50396	20,244.06	20,456.33
-0.75	-0.75	0.50344	20,239.83	20,452.10
-0.75	-0.65	0.50301	20,236.21	20,448.48
-0.75	-0.55	0.50267	20,233.16	20,445.44
-0.75	-0.45	0.50241	20,230.68	20,442.95
-0.75	-0.35	0.50222	20,228.74	20,441.01
-0.75	-0.25	0.50211	20,227.32	20,439.59
-0.75	**-0.15**	**0.50203**	**20,226.41**	**20,438.68**
-0.75	-0.05	0.50209	20,226.01	20,438.28
-0.75	$+0.05$	0.50218	20,226.10	20,438.37
-0.75	$+0.15$	0.50233	20,226.67	20,438.95
-0.75	$+0.25$	0.50255	20,227.73	20,440.01
-0.75	$+0.35$	0.50283	20,229.27	20,441.55

Table 11. Predictive accuracy measures and model selection measures of the BivGEV models with Frank copula for d values of (τ_1, τ_2). In bold the selected model

Credit bureau default	Default P2P	Predictive Accuracy	Model selection measures	
τ_1	τ_2	MSE_+	AIC	BIC
-0.75	-0.85	0.50394	20,243.99	20,456.26
-0.75	-0.75	0.50342	20,239.76	20,452.04
-0.75	-0.65	0.50299	20,236.15	20,448.42
-0.75	-0.55	0.50265	20,233.11	20,445.38
-0.75	-0.45	0.50239	20,230.64	20,442.91
-0.75	-0.35	0.50220	20,228.70	20,440.98
-0.75	-0.25	0.50209	20,227.29	20,439.57
-0.75	**-0.15**	**0.50204**	**20,226.47**	**20,438.71**
-0.75	-0.05	0.50207	20,226.00	20,438.28
-0.75	$+0.05$	0.50216	20,226.10	20,438.38
-0.75	$+0.15$	0.50231	20,226.69	20,438.96
-0.75	$+0.25$	0.50253	20,227.76	20,440.03
-0.75	$+0.35$	0.50281	20,229.31	20,441.58

Table 12. Predictive accuracy measures and model selection measures of the BivGEV models with Joe copula for differen of (τ_1, τ_2). In bold the selected model

Credit bureau default	Default P2P	Predictive Accuracy	Model selection measures	
τ_1	τ_2	MSE_+	AIC	BIC
-0.75	-0.85	0.50393	20,244.03	20,456.3
-0.75	-0.75	0.50342	20,239.85	20,452.1
-0.75	-0.65	0.50299	20,236.26	20,448.5
-0.75	-0.55	0.50265	20,233.24	20,445.5
-0.75	-0.45	0.50239	20,230.78	20,443.1
-0.75	-0.35	0.50221	20,228.86	20,441.1
-0.75	-0.25	0.50209	20,227.46	20,439.7
-0.75	**-0.15**	**0.50205**	**20,226.57**	**20,438.8**
-0.75	-0.05	0.50207	20,226.19	20,438.5
-0.75	$+0.05$	0.50216	20,226.30	20,438.6
-0.75	$+0.15$	0.50232	20,226.89	20,439.2
-0.75	$+0.25$	0.50254	20,227.97	20,440.2
-0.75	$+0.35$	0.50282	20,229.53	20,441.8